DOI: 10.14393/DLv17a2023-21
e1721

Artigo

Abordagem baseada em Aumento de Dados para Avaliação Automática de Leiturabilidade

A Data Augmentation approach to Automated Readability Assessment

Luiza Cunha de MENEZES[*]
Aline PAES[**]
Maria José Bocorny FINATTO[***]

RESUMO: Embora estudos sobre como medir a leiturabilidade de um texto remontem ao século passado, ainda não há um consenso sobre quais seriam as melhores métricas. Ferramentas de Processamento de Linguagem Natural (PLN) podem apoiar esta tarefa, mas dependem de um grande número de amostras para treinamento, o que é uma barreira para seu avanço. O objetivo principal deste artigo é analisar o impacto de determinados métodos de aumento de dados (AD) para enfrentar essa barreira e apoiar a classificação de leiturabilidade no português brasileiro (PB). Para tanto, foi estabelecido um *corpus* pareado e classificado, com textos originais complexos e suas versões simplificadas sobre temas de Ciências, desenvolvido por linguistas. Esse *corpus* foi aumentado com técnicas agnósticas de AD: substituição por sinônimos (SS) e retrotradução (RT). Foram avaliados 75 modelos com diferentes técnicas e combinações de atributos de entrada. O melhor resultado obtido para o conjunto dos textos do *corpus* sem aumento foi de 94,0% de taxa de acerto. Este resultado subiu para 95,2% combinando-se as métricas do sistema NILC-Metrix com representações vetoriais de texto contextualizadas. Quando comparados a outros trabalhos voltados para o PB, a metodologia proposta gerou um aumento na taxa de acerto em um domínio distinto ao de treino. Conclui-se que o modelo treinado com AD demonstra capacidade igual ou superior àqueles treinados sem aumento e, ao mesmo tempo, apresenta maior generalização quando aplicado a outros domínios.
PALAVRAS-CHAVE: Processamento de Linguagem Natural. Substituição por Sinônimo. Retrotradução. Aumento de Dados. Avaliação Automática de Leiturabilidade.

ABSTRACT: Studies about how to measure text readability reassemble the last century. Nonetheless, there is no consensus on which could be the best metrics. Tools regarding the field of Natural Language Processing (NLP) may support this task but are dependent on a high number of samples for training, and that is a bottleneck to its advancement. The main goal of this paper is to analyze the impact of a couple of data augmentation (DA) methods to support the readability classification task in Brazilian Portuguese (BP) to mitigate the bottleneck problem. In this sense, we worked on a paired and classified *corpus* created by linguists. The *corpus* is about science, and each text contemplates its original and simplified

[*] Universidade Federal Fluminense (UFF) - lumenezes@id.uff.br
[**] Universidade Federal Fluminense (UFF) - alinepaes@ic.uff.br
[***] Universidade Federal do Rio Grande do Sul (UFRGS) - mariafinatto@gmail.com

versions. About the methodology, we considered two agnostic tasks: synonym replacement and back-translation and evaluated 75 models with different techniques and combinations of input features. For the trained model with the *corpus* without DA, the best score reached 94.0% of the hit rate. When combining the NILC-Metrix metrics and contextualized word embeddings, the results overtook 95.2%. Compared to other papers applied to the BP, the proposed methodology improved the hit rate considering a distinct training domain. Our results demonstrate that the capacity of DA methods can be equal to or greater than those trained without augmentation and, at the same time, present greater generalization when applied to other domains.
KEYWORDS: Natural Language Processing. Synonym Replacement. Back-translation. Data Augmentation. Automatic Readability Assessment.

Artigo recebido em: 13.02.2023
Artigo aprovado em: 03.04.2023

1 Introdução

A comunicação é um elemento essencial na interação social humana. Para que ocorra de forma fluida e eficiente, é importante que locutor e interlocutor compartilhem do mesmo código, convenções e valores associados à linguagem. No entanto, as variações, tanto do código em si, como de interpretações destes códigos e valores, podem impedir que a comunicação aconteça de modo efetivo.

No que diz respeito à comunicação textual via escrita, uma forma de se estimar se um determinado conteúdo estaria de acordo com o que se propõe informar, *i.e.*, se estaria de acordo com a habilidade interpretativa do leitor-destinatário, é a quantificação de elementos associados à propriedade denominada leiturabilidade[1]. Todavia, atribuir um número à característica subjetiva da leiturabilidade tem sido motivo de estudos ao longo das últimas décadas, em diferentes idiomas. Estudos iniciais, voltados para desenvolver textos em inglês escrito em versão facilitada (*plain language*) já consideraram combinações envolvendo métricas de análise linguística e psicolinguística (KINCAID *et al.*, 1967; KLARE *et al.*, 1963; CAYLOR *et al.*, 1973;

[1] De acordo com Ponomarenko (2018), leiturabilidade é uma condição de facilidade de leitura criada por escolhas de conteúdo, estilo, design e organização que se alinham ao conhecimento prévio, escolaridade, interesse e motivação do público leitor.

MARTINS *et al*., 1996). Entretanto, até o momento, não há consenso entre os estudiosos da área da Linguística descritiva ou computacional em termos de estabelecer uma categorização ou escala numérica ideal para estimar a leiturabilidade textual. Isto porque as métricas existentes são bastante superficiais e se relacionam às propriedades estruturais dos textos; além disso, tendem a não considerar elementos como a interação entre texto e leitor (SANTOS, 2010; SCARTON; ALUÍSIO, 2010; FINATTO, 2020).

Neste sentido, torna-se interessante avaliar o uso de abordagens contextualizadas baseadas em métodos de Processamento de Linguagem Natural (PLN). Esta é uma área na interseção entre a Linguística e a Inteligência Artificial (IA) que visa desenvolver métodos computacionais para analisar ou gerar textos de modo automático. Com o advento do Aprendizado de Máquina (AM), a área de PLN tem abordado o uso de dados para gerar modelos que resolvam alguma tarefa linguística. Por exemplo, no contexto de análise de leiturabilidade, a tarefa poderia ser a de classificar textos como mais ou menos complexos, em um dado contexto, e o modelo seria uma abstração matemática que fizesse tal classificação. Tal abstração seria automaticamente induzida a partir de um conjunto de textos previamente classificados.

Assim, com os avanços em PLN, em termos de um enfoque computacional para as métricas supracitadas, a tarefa de avaliação estimativa de leiturabilidade é denominada por *Automatic Readability Assessment (ARA)*, ou Avaliação Automática de Leiturabilidade (AAL), em português. Neste contexto, pesquisadores que atuam em Linguística Computacional investigam o uso de modelos de linguagem estatísticos para capturar características linguísticas e desenvolver modelos de AM para classificação automatizada de textos quanto à sua leiturabilidade. Isso é feito considerando-se a identificação e quantificação de elementos da superfície dos textos escritos, diferentes cenários de comunicação e variados perfis de leitores.

Observa-se, nessa trajetória, que a realização automática de tarefas relacionadas ao domínio da comunicação e usos da linguagem é intrinsecamente complexa.

Estendem-se a estes desafios questões relacionadas à disponibilidade, geralmente escassa, de conjuntos robustos de dados linguísticos, coerentes com o problema analisado. Isto porque, conforme apontado por Brill (2003) e Sahin (2022), uma das maiores barreiras ao PLN avançado é o gargalo de aquisição automática de conhecimento linguístico via *corpora* textuais. Afinal, os modelos atuais de PLN dependem de um número elevado de amostras no conjunto de dados de treinamento, pois, sem dados robustos, eles não são capazes de generalizar a tarefa além dos dados de treinamento. Nesses dados, uma necessidade é a anotação dos *corpora*, que precisam trazer uma camada de informação linguística sobre as palavras que os perfazem, como informações semânticas e sintáticas, por exemplo.

Visando melhorar o desempenho de sistemas automáticos de PLN frente à carência de grandes conjuntos de dados provenientes de usos de linguagem, pesquisadores introduziram métodos de Aumento de Dados (AD) no domínio de análise e processamento textual, com a finalidade de aumentar o tamanho de uma dada amostra a ser utilizada em um treinamento.

Dado este contexto, o trabalho aqui relatado busca uma alternativa para aumentar a quantidade de amostras a partir de um *corpus* classificado em termos de graus de leiturabilidade. Isso é feito por meio de dois métodos de AD pertencentes ao domínio de PLN, que são a Substituição por Sinônimo (SS) e a Retrotradução (RT). Ambos visam aliviar os impactos gerados pelo gargalo de dados apontados por diversos autores (CAO *et al*., 2020; BRILL, 2003; SAHIN, 2022; FREITAS, 2022) e que são ampliados pelos contextos linguístico e comunicativo apresentados por Finatto (2020) e Freitas (2022). Nesse sentido, a hipótese de pesquisa deste trabalho é que o uso de determinados métodos de AD em PLN potencializa o treinamento dos modelos linguísticos computacionais relacionados à classificação binária de complexidade textual em português Brasileiro (PB).

No restante deste artigo, conceitualizamos a área de estudo, apresentamos trabalhos relacionados, a metodologia proposta, seus resultados e conclusões.

2 Pressupostos teóricos

Este artigo relata uma investigação sobre a tarefa de AAL, a partir de modelos de AM e PLN, treinados a partir de uma amostra de dados linguísticos aumentados automaticamente. A tarefa é modelada como um classificador automático, treinado a partir de um conjunto de textos anotados com suas classes correspondentes, seguindo a metodologia de aprendizado de máquina supervisionado (IMPERIAL, 2021). Conforme explicado por Freitas (2022), o objetivo do AM é resolver problemas sem que os modelos tenham sido explicitamente programados para isto. Desta forma, padrões para a resolução das tarefas são aprendidos automaticamente por meio de exemplos, chamados de dados de treino. No caso de um classificador textual binário em termos de leiturabilidade, o objetivo da tarefa é classificar um texto como simples ou não.

2.1. Representações de Atributos Textuais

Devido à natureza subjetiva e complexa da linguagem, que se manifesta em alta dimensionalidade e esparsidade (AGGARWAL, 2018), torna-se um desafio traduzir palavras em unidades discretas e estáveis (FREITAS, 2022). Assim, uma das formas de comprimir dados e facilitar a identificação de padrões pela máquina é por meio de técnicas de representação de atributos textuais (AGGARWAL, 2018). Neste sentido, exploraram-se, neste artigo, duas abordagens computacionais de representação de atributos textuais:

- A **incorporação de palavras**, ou, em inglês, *word embeddings*, permite capturar características semântico-lexicais de uma linguagem natural e transportá-las para um espaço latente[2]. Uma *word embedding* pode ser entendida como uma representação das palavras de um texto de forma numérica (AGGARWAL,

[2] Um espaço latente pode ser definido como um espaço abstrato multidimensional que codifica uma representação interna significativa de eventos observados externamente (CHAQUET-ULLDEMOLINS *et al*., 2022)

2018; FREITAS, 2022), de modo que palavras semanticamente semelhantes são representadas vetorialmente mais próximas umas das outras (FREITAS, 2022; AGGARWAL, 2018), seguindo a semântica distribucional.

A incorporação de palavras pode ser de forma estática, em que as palavras sempre terão a mesma representação vetorial, independente do contexto, ou de forma contextualizada, em que a representação de uma palavra dependerá do contexto corrente. Nesse caso, os modelos conhecidos por *Transformers* processam uma sequência de palavras de uma só vez e mapeiam as dependências relevantes entre as palavras, independente de quão distantes as palavras apareçam no texto (VASWANI *et al*., 2017). Conforme apontado por Chuchu (2022), os *Transformers* são o estado da arte em modelos neurais de linguagem. A arquitetura do *Transformer* é uma rede neural composta por camadas de entrada, que transformam os dados de entrada em uma representação vetorial; e camadas de saída, capazes de transformar essas representações em saídas legíveis. Entre essas camadas, existem camadas de atenção, que calculam a importância de cada elemento em relação aos outros elementos do conjunto de dados, capturando as dependências entre as palavras. Um ponto de atenção é que estas redes neurais são computacionalmente intensas e precisam de muitos recursos para treino. Isto significa que um número elevado de *tokens*[3] processados ao mesmo tempo podem levar a resultados imprecisos ou impraticáveis (SOUZA *et al*., 2020). Por isso, essas

[3] Segundo Freitas (2022), a tokenização é o processo de dividir um texto em unidades menores chamadas *tokens*. Tais unidades podem ser palavras, sinais de pontuação ou símbolos. Freitas (2022) aponta para três estratégias de tokenização: (1) baseada em palavra, que divide o texto em palavras gráficas, ou seja, o que aparece entre espaços em branco ou sinais de pontuação; (2) baseada em caracteres, que cria uma representação numérica para cada letra; e (3) baseada em subpalavra, uma abordagem intermediária, que divide as palavras em pedaços menores mais informativos do que os caracteres, mas com menos caracteres do que as palavras gráficas. Ainda que conceitualmente palavra e *token* tenham significados diferentes, a primeira sendo uma unidade linguística e a segunda, uma unidade textual, dado o presente contexto de vetorização, ambos os termos serão utilizados de forma intercambiável neste trabalho.

abordagens são projetadas para lidar com textos de tamanho moderado, e tendem a ter restrições no número de *tokens* a serem processados.

No caso do PB, o BERTimbau é um modelo BERT, ou seja, um codificador *Transformer* que trabalha com representações de forma bidirecional[4], pré-treinado especificamente para o PB (SOUZA *et al*., 2020). Este modelo está integrado à biblioteca de código aberto *transformers*, disponibilizada pelo grupo *Hugging Face* (WOLF *et al*., 2020).

- As **métricas de análise linguística e psicolinguística** são embasadas em medidas textuais de coesão, coerência, nível de complexidade, clareza, precisão e consistência. Por isso, podem ser consideradas uma interessante representação numérica da estrutura de um texto em termos de estimar a sua leiturabilidade, haja vista que cada uma dessas métricas medem aspectos da qualidade de um texto e podem ser usadas para avaliar textos em diferentes domínios. Atualmente, a ferramenta NILC Metrix (LEAL *et al*., 2021; LEAL *et al*., 2022a) é um dos recursos mais completos para gerar métricas de análise linguística e psicolinguística para o PB.

Para cada uma das possibilidades de combinações das representações de atributos acima, é possível treinar e testar classificadores com diferentes entradas para textos originais e aumentados. Assim, quantificam-se os impactos das inclusões de dados sintéticos em termos de classificação binária de complexidade textual para o PB.

2.2 Contextualização das Técnicas de Aumento de Dados

A abordagem de AD tem por objetivo o aumento da variedade de um conjunto inicial de dados sem a necessidade de nova coleta e categorização. O AD deve fornecer

[4] Bidirecional neste sentido significa que uma janela de palavras, tanto à esquerda quanto à direita da palavra que se deseja representar de forma numérica, é levada em consideração pelo modelo.

uma alternativa para obtenção de mais dados, de modo que, um método ideal deve ser fácil de implementar e capaz de melhorar o desempenho do modelo (FENG *et al*., 2021).

Os métodos de AD tiveram sua origem no campo de dados em forma de imagem (BAYER *et al*., 2021). Com imagens, transformações simples como cortar, girar e variar suas cores são úteis, pois permitem criar variações e particularidades que facilitam a generalização dos modelos de AM. Isto significa que as operações aplicadas às imagens normalmente não alteram a natureza daquilo que foi capturado, apenas as reapresentam em formatos diferentes. No domínio do PLN, a geração de exemplos aumentados de textos escritos, capazes de capturar as invariâncias e alternâncias linguísticas desejadas, é uma tarefa bem menos óbvia (FENG *et al*., 2021; TAYNAN; COSTA, 2020). Ao ponderarmos alterações análogas às supracitadas para imagens, em um texto, as estratégias intuitivas seriam a remoção de trechos em sentenças, reordenação ou inserção aleatória de palavras. Entretanto, tais mudanças implicam na possível violação de uma série de regras léxicas e sintáticas, além do distanciamento do conteúdo inicial em termos do significado de palavras e do significado do todo do texto. No entanto, conforme apontado por Feng *et al*. (2021), é possível se inspirar nos métodos de AD em imagens para textos. Por exemplo, aplicar uma escala de cinza em uma imagem colorida pode ser entendida como uma atenuação de aspectos linguísticos, como uma mudança do grau superlativo dos adjetivos contidos no texto.

Especificamente sobre o domínio de PLN, Bayer *et al*. (2021) aponta que pesquisas conduzidas para o AD são recentes e eram escassas até 2019. Neste contexto, destaca-se o fato de que há uma lacuna teórica que fundamente os estudos em AD; segundo Feng *et al*. (2021), a maioria dos estudos pode mostrar empiricamente que uma técnica de AD funciona, mas é um desafio medir a qualidade de uma técnica sem recorrer a um experimento em grande escala.

Assim, o uso de métodos de AD ainda apresenta resultados limitados em termos de ganhos de desempenho, e como consequência, esses métodos automáticos

têm sido pouco explorados pela comunidade. Isso vem ocasionando, por exemplo, uma carência de compreensão entre AD e implicações no modelo de aprendizado (TAYNAN; COSTA, 2020). Enquanto isso, o estudo de Longpre *et al*. (2020) levanta a hipótese de que os métodos de AD aplicados ao PLN só podem ser benéficos e eficientes se introduzirem novos padrões linguísticos. Desta forma, os autores sugerem que, ao considerar grandes modelos de língua pré treinados, muitos métodos de AD não conseguem gerar maiores ganhos, pois tais modelos já são invariantes a diversas transformações.

Por isso, neste artigo, investigaremos a hipótese de que o uso de determinados métodos de AD em PLN potencializa o treinamento dos modelos linguísticos no contexto de complexidade textual, especificamente na avaliação quantitativa da tarefa de AAL.

3 Trabalhos Relacionados

No que diz respeito ao estado-da-arte em AAL, destacam-se os modelos propostos por Martinc *et al*. (2021) e Lee *et al*. (2021), pois foram capazes de aumentar a precisão da classificação em um *corpus* de língua inglesa. Esses autores apresentam o *corpus* Weebit (FENG *et al*., 2009) e um modelo computacional de língua que amplia em cerca de 4% a precisão por meio da incorporação de atributos contextualizados.

De acordo com Lee *et al*. (2021), este resultado de incremento sugeriu, de forma inédita, que modelos de redes neurais com atributos gerados automaticamente por transferência de aprendizado podem ser mais eficazes do que os modelos de AM tradicionais na tarefa de AAL. No trabalho de Lee *et al*. (2021), além do *corpus Weebit*, são considerados outros dois *corpora* da língua inglesa, *OneStopEnglish* e *Cambridge* (XIA *et al*., 2019). Os autores exploram o uso de um modelo híbrido, em que são combinados os resultados das previsões de um classificador por transferência de aprendizado, *i.e.*, por meio de uma rede neural considerando um modelo de incorporação de palavras pré-treinado, e atributos linguísticos envolvidos por um

modelo não-neural, como um regressor logístico (RL). Em comparação com o estudo anterior de Martinc *et al*. (2021), Lee *et al*. (2021) conseguiram um aumento de aproximadamente 20%, atingindo uma taxa de acertos (acurácia) de 99% para o *corpus OneStopEnglish*. Este valor notório é apontado como uma preocupação de *overfitting*[5] para aquele domínio textual.

Assim sendo, segundo os autores antes citados, sem uma metodologia capaz de conectar vários conjuntos de dados ou um novo grande conjunto público de dados para AAL, será sempre um desafio desenvolver um modelo computacional de língua de uso geral. Adicionalmente, uma conclusão importante na análise de Lee *et al*. (2021) é que conjuntos menores de dados se beneficiam mais do uso de atributos linguísticos do que da incorporação de palavras. Apenas para fins comparativos, o *corpus* Weebit contém 3.125 textos, o OneStopEnglish, 567, e o Cambridge, 331.

Ainda que os primeiros trabalhos de AAL (SI; CALLAN, 2001; SCHWARM; OSTENDORF, 2005) já tenham pouco mais de duas décadas de existência, trabalhos voltados para o PB foram fomentados a partir do surgimento do Coh-Metrix-PORT (MCNAMARA; LOUWERSE; GRAESSER, 2002). No entanto, conforme apontado por Scarton *et al*. (2010a), até o ano de 2010, a única ferramenta a considerar abordagens em PLN para avaliação de leiturabilidade no PB havia sido realizada pelo trabalho de Scarton *et al*. (2010a). No que diz respeito à aplicação da tarefa de AAL para trabalhos voltados especificamente para o PB, identificaram-se:

- Scarton e Aluísio (2010) apresentam aplicações de PLN para a análise da inteligibilidade de textos. Dentre as aplicações apresentadas, é considerada a

[5] Conforme apontado por Freitas (2022), o *overfitting* pode ser compreendido com a analogia entre um modelo de aprendizado de máquina e um estudante. Se o estudante tem acesso prévio aos resultados de uma prova, ele acerta as questões não porque aprendeu o conteúdo, mas sim porque decorou as respostas. De forma análoga, se os dados de treino usados para um modelo automático de classificação se aproximarem muito dos dados de teste, o modelo acertará as classes do teste apenas por estar copiando exatamente o que observou no treino.

criação de classificadores binários entre textos complexos e simples. Para isto, foi desenvolvido um *corpus* combinado, formado por textos classificados conforme o público-alvo da revista ou jornal de origem. O melhor classificador treinado atingiu em média um F-*score*[6] de 0,97, considerando as métricas do Coh-Metrix-Port combinadas com o índice Flesch como sendo os atributos de entrada do modelo.

- Aluísio *et al*. (2010) apresentam um recurso incorporado à ferramenta Simplifica (SCARTON *et al*., 2010b), cujo intuito é o de categorizar textos em três níveis de leiturabilidade: (1) textos originais voltados para leitores avançados, (2) textos naturalmente simplificados voltados para pessoas com nível de alfabetização básico e (3) textos fortemente simplificados para pessoas com nível de alfabetização rudimentar. Como atributos do modelo classificatório, foram considerados três grupos de atributos: o primeiro contém atributos derivados da ferramenta Coh-Metrix-Port, o segundo contém características que refletem a incidência de construções sintáticas particulares desenvolvidas pelos próprios autores, e o terceiro contém características derivadas de um modelo de língua desenvolvido com base na ferramenta SRILM (STOLCKE, 2002), produzido com artigos advindos da Folha de São Paulo. O modelo classificatório com melhor resultado foi aquele que considerou a combinação dos três grupos de atributos e alcançou um F-*Score* de 0,913 para o nível (1), 0,483 para (2) e 0,732 para (3).
- Scarton *et al*. (2010a) consideram duas classes para distinção dos textos entre simples (para leitores entre 7 e 14 anos) e complexo (para adultos). Nos testes, avaliou-se o impacto de diferentes gêneros e domínios por meio da criação de dois classificadores independentes, um treinado apenas com textos de notícias

[6] O *F-score* é uma métrica calculada pela média harmônica entre precisão (taxa de verdadeiros positivos em relação ao total de amostras de positivos) e revocação (taxa de verdadeiros negativos em relação ao total de amostras de negativos), e varia de 0 a 1.

e outro treinado apenas com textos de divulgação científica. Adicionalmente, experimentaram-se algoritmos de seleção de atributos, a fim de selecionar aqueles mais relevantes de um conjunto extraído do Coh-Metrix Port. Os experimentos mostraram que o algoritmo com todas as métricas do Coh-Metrix Port obteve o melhor desempenho e que o classificador treinado em textos de jornal foi capaz de generalizar e classificar consideravelmente bem os textos de divulgação científica.

- Wilkens *et al*. (2016) fizeram uso da iniciativa *Web as Corpus* (WaC)[7] como uma ferramenta para geração de grandes *corpora* classificados em termos de leiturabilidade. Para anotação dos *corpora* coletados, os autores utilizaram um classificador cujos atributos de entrada eram sete métricas de coesão e coerência textual. O treino foi realizado com um *corpus* mais controlado, o Wikilivros[8], que é separado em três níveis conforme o sistema educacional brasileiro: 33 livros do Ensino Fundamental; 65 livros do Ensino Médio e 21 livros de cursos de graduação. O classificador atingiu em média um F-*score* de 0,691, com precisão de 0,702 e revocação de 0,689.
- Hartmann e Aluísio (2020) investigaram um método de adaptação lexical em textos para o Ensino Fundamental. Ainda que este trabalho não esteja direcionado diretamente para a tarefa de AAL, ao investigar etapas do processo de adaptação textual, são consideradas abordagens para a identificação de palavras complexas. No processo de classificação de palavras em simples ou complexas, Hartmann e Aluísio (2020) avaliaram desde abordagens clássicas até as mais modernas, incluindo uso de *embeddings* contextualizados. O uso do *embedding* oriundo do modelo Elmo apresentou melhor desempenho, com acurácia média de 97,7%.

[7] WaC é um conjunto de ferramentas que permite o acesso à World Wide Web como um *corpus.*
[8] Disponível em: https://pt.wikibooks.org. Acesso em: 20 fev. 2023.

- Leal (2019) avaliou métodos de predição de complexidade de frases para o PB. Para isto, foram criados dois *corpora*, um com sentenças alinhadas pelo PorSimples, o PorSimplesSent (LEAL *et al*., 2018), e outro com métricas de rastreamento ocular e normas de previsibilidade para estudantes de nível superior, denominado RastrOS (LEAL *et al*., 2022b). Foi considerada a versão mais recente da ferramenta NILC-Metrix (LEAL *et al*., 2021) (com 200 métricas), bem como abordagens de transferência de aprendizado com adição das métricas de rastreamento ocular. Leal (2019) atingiu o nível do estado-da-arte para a tarefa de predição da complexidade de frases no PB, com 97,5% de acurácia. Com base no melhor método desenvolvido, o autor criou a aplicação Simpligo (LEAL *et al*., 2019; LEAL, 2020), que atribui um índice de complexidade individual para a sentença informada.

Certamente, o grupo de pesquisa Núcleo Interinstitucional de Linguística Computacional (NILC) é uma referência no uso de PLN avançado em diversas tarefas, incluindo a de simplificação textual e de análises para níveis linguísticos variados, dedicando-se especialmente ao PB. No contexto de trabalhos sobre leiturabilidade, destaca-se, no NILC, o projeto PorSimples (ALUÍSIO *et al*., 2008). Mais recentemente, o NILC desenvolveu o trabalho de Leal (2019) sobre medidas de complexidade textual, o que culminou na criação da ferramenta Simpligo (LEAL *et al*., 2019). O Simpligo encontra-se disponível para uso *on-line* na plataforma do NILC (LEAL, 2020).

O projeto PorSimples, na sua origem, teve por objetivo apoiar a facilitação de informações para crianças e adultos em processo de alfabetização ou para pessoas com algum tipo de deficiência ou condição associada com a leitura de materiais escritos. Ao longo da trajetória do PorSimples, foram criadas diferentes ferramentas computacionais associadas à simplificação de textos, como o Facilita (WATANABE *et al*., 2009) e o Simplifica (SCARTON *et al*., 2010b). Este é um editor destinado a produtores de conteúdo que desejam criar textos simplificados adequados ao mesmo

público, e aquele é um plug-in de navegador para simplificar automaticamente textos de *sites*.

No entanto, a tarefa aqui proposta se diferencia de tais trabalhos ao se concentrar na classificação textual para estimar a leiturabilidade e no uso de ferramentas que sejam capazes de mitigar o gargalo da falta de grandes quantidades de dados linguísticos. Assim, aproxima-se mais do trabalho proposto por Wilkens *et al*. (2016). O trabalho de Wilkens *et al*. (2016) é particularmente interessante, haja vista que os autores visam a classificação por leiturabilidade e sugerem o uso da iniciativa *WaC* para reduzir o já citado gargalo linguístico.

Finalmente, em trabalhos que lidam com outros idiomas com menos recursos do que a língua inglesa, como é o nosso caso com o português, destaca-se o estudo feito por Imperial (2021) para o idioma filipino. Segundo o autor, evidências de eficácia de métodos baseados em transferência de aprendizado, como no trabalho de Martinc *et al*. (2021), valem apenas para conjuntos de dados com alta volumetria.

Assim, Imperial (2021) treinou classificadores a partir de *embeddings* oriundos do modelo BERT e métricas linguísticas concatenados. Para o *corpus Adarna House*, que contém 265 livros em filipino, o uso da concatenação de *embeddings* e de métricas linguísticas superou o uso individual destes conjuntos de atributos, alcançando um valor de medida F de 0,571.

4 Metodologia baseada em AD para AAL

Este artigo considera tarefas agnósticas[9] de AD devido à facilidade de implementação e menor custo computacional. Quanto aos métodos de AD selecionados, é importante entender a sua escolha conforme o contexto deste trabalho. Esse contexto pode ser resumido em duas premissas:

[9] Tarefas que são generalizáveis e consequentemente, não são cunhadas para nenhuma tarefa em particular (LONGPRE *et al*., 2020).

- Os recursos de PLN disponíveis para a língua portuguesa (ainda que não tão abrangentes quanto da língua inglesa) permitem o uso de mecanismos para substituições e conversões interlinguísticas;
- Conforme Finatto (2020), a condição de leiturabilidade de um texto é multifatorial. Está fortemente atrelada ao tema e ao estilo do texto, bem como ao uso de terminologias, de vocabulário mais ou menos frequente, tipo de organização sintática ou de tipos de frases, entre outros tantos elementos.

Sendo assim, métodos de AD que envolvessem maior perda potencial de significado (seja por redução de contexto ou inclusão de ruídos) foram desconsiderados nesta análise. Com esta restrição, foram selecionados e desenvolvidos métodos nas seguintes categorias:

1. **Substituição por Sinônimo (SS)**: é um método bastante popular de AD que parafraseia instâncias de texto substituindo certas palavras por sinônimos. Esta tarefa se aproxima da Simplificação Léxica (SL), no entanto, a SL pertence à área de adaptação textual e tem por finalidade reduzir a complexidade lexical ou sintática um texto, preservando seu significado (HARTMANN; ALUÍSIO, 2020). Conforme pontuado por Wilkens *et al*. (2014), a SL tem por objetivo substituir palavras complexas por sinônimos ou palavras semanticamente próximas, que sejam de mais fácil compreensão. Logo, a diferença entre SS e SL reside no fato de que, em termos semânticos, é possível alterar palavras e/ou expressões por um outro conjunto de palavras que não necessariamente estabeleçam uma relação de sinonímia entre si.
2. **Retrotradução (RT)**, ou do inglês *back-translation*: esta abordagem é particularmente interessante porque possui alta capacidade de parafraseamento, permitindo alterações tanto léxicas quanto sintáticas (BAYER *et al*., 2021). Segundo Bayer *et al*. (2021), por se utilizar a tarefa de tradução do

texto, o conteúdo é preservado e apenas as características estilísticas baseadas nos traços do autor são excluídas ou alteradas.

4.1 Métodos Propostos para o Aumento de Dados para AAL

A seguir, explica-se como foram desenvolvidos os métodos de AD propostos neste estudo, antes citados.

4.1.1 Substituição por Sinônimo (SS)

Para a execução desta tarefa de substituição, foi necessário inicialmente criar um contexto de palavras de modo a identificar aquelas que estabelecessem uma relação de sinonímia entre si. Para isto, partiu-se da definição de um universo de palavras que são simples, diferentemente do formato de substituição apresentado por Wilkens *et al*. (2014), em que os autores definem como primeira etapa a definição de palavras consideradas difíceis para serem substituídas.

Para construir este contexto de palavras, foram considerados os seguintes conjuntos de dados disponíveis para o PB:

- Tep 2.0 (DIAS-DA-SILVA; MORAES, 2003): uma coleção de palavras do PB e respectivos sinônimos, agrupadas em conjuntos (USP, 2022b); no inglês, este tipo de coleção é conhecido por *synset*. Naturalmente, uma palavra pode ter diversos sinônimos a depender do contexto. Assim, o Tep 2.0 traz os contextos expressos por meio de um número sequencial.
- *Wordlist* do CorPop (PASQUALINI, 2018): uma lista de palavras com seus respectivos números de frequência, obtidos por um compilado de textos do PB popular escrito e selecionados com base no nível de letramento médio do país.
- PortiLexicon (USP, 2022a): contém mais de 1,2 milhão de formas de palavras em português com suas respectivas informações morfológicas e morfossintáticas, seguindo o modelo internacional de Dependências Universais. O léxico é

baseado no Unitex PB (MUNIZ, 2004) e faz parte do projeto POeTiSA (USP SÃO CARLOS, 2021).

Estes conjuntos foram combinados de modo que foi gerado um léxico complementar (ou contexto de palavras), composto por 12.564 palavras, dentre verbos, adjetivos, advérbios e substantivos, com informações de (a) sinônimo mais simples, (b) forma no infinitivo, (c) gênero e (d) número.

A metodologia para construção deste léxico de itens e sinônimos considerou a combinação de cada palavra e seu conjunto de sinônimos apresentados no TeP 2.0, em suas formas no infinitivo (advindas da base léxica - PortiLexicon), com a frequência de cada palavra correspondente no CorPop, também em sua forma no infinitivo dada pelo PortiLexicon. Assim, o sinônimo com maior número de hits do CorPop, *i.e*. maior frequência, foi eleito como sinônimo mais simples. Observa-se, portanto, que o CorPop apresenta um papel de extrema importância para o formato de SS aqui apresentado, haja vista que:

- Foi desenvolvido a partir da análise de dados sobre o nível de letramento dos leitores brasileiros e das características que poderiam compor um padrão de simplicidade textual em um *corpus* adequado a estes leitores (PASQUALINI, 2018), estando diretamente relacionado aos objetivos motivadores desta pesquisa;
- Considera dados de frequência para cada uma das palavras consideradas mais simples ou comuns no vocabulário popular brasileiro[10].

Na sequência, foram incluídas informações de gênero e número das palavras a partir do PortiLexicon para facilitar a SS, garantindo coesão e coerência textual. As

[10] Conforme um dos principais pontos de análise apresentados por Wilkens *et al.* (2014), ao contrário do que geralmente se supõe, o tamanho de uma palavra não é um bom indicativo para categorização de palavras simples e complexas, mas sim, a frequência de sua utilização.

substituições automaticamente realizadas nesta metodologia consideraram a substituição de uma palavra flexionada em um determinado gênero e grau por outra na mesma flexão. Em casos de impossibilidade sintática, optou-se pela não substituição. Ainda que este comportamento restrinja o número de substituições possíveis, a redução na complexidade do processo de substituição é significativa. Isto porque uma mudança de flexão em uma única palavra pode implicar na alteração de flexão em diversas palavras dependentes em diferentes sentenças do texto.

Assim sendo, a partir do léxico criado, é feita a varredura de cada um dos textos contidos no *corpus* de treinamento, em que foram verificadas todas as palavras dos textos por meio do *tokenizador* da biblioteca NLTK (BIRD *et al*., 2009). Apenas os tokens identificados no *corpus* e que não fossem nomes próprios dados pela biblioteca Spacy (HONNIBAL; MONTANI, 2017) são passíveis de serem substituídos por seus respectivos sinônimos.

Devido ao fato de existirem diferentes contextos de sinônimos expressos pelo Tep 2.0, uma mesma palavra pode ser substituída por mais de um sinônimo. Nesse sentido, a metodologia proposta considerou uma seleção inspirada no algoritmo genético da roleta[11] (SHUKLA *et al*., 2015). É importante ressaltar que, por se tratar de uma estratégia usada para AD, é essencial que a SS desenvolvida seja capaz de manter o rótulo inicial do texto, *i.e.* simples ou complexo. Portanto, no caso de um texto originalmente simples, os sinônimos com maior frequência no CorPop possuíam uma porção maior da roleta, e aqueles com frequência mais baixa possuíam uma porção relativamente menor. O processo inverso válido foi realizado para os textos originalmente complexos.

[11] Os algoritmos genéticos recebem este nome porque são inspirados na teoria da evolução. No caso do algoritmo da roleta, é inspirado na teoria da seleção natural, em que os indivíduos mais adaptativos são selecionados com maior probabilidade para reprodução.

4.1.2 Retrotradução (RT)

Esse método visa gerar, automaticamente, sentenças com o mesmo significado e anotação das sentenças originais por meio da tradução de uma língua X para uma língua Y, com uma retrotradução para a língua X. O processo em si de construção da tarefa é simples, visto que a biblioteca *Deep Translator* (BACCOURI, 2020) permite a parametrização do idioma de origem para o idioma desejado e não há restrição no número de palavras.

No entanto, há uma preocupação na escolha do idioma intermediário, *i.e.*, da língua Y, porque afeta diretamente o resultado do processo de RT e, consequentemente, a manutenção do rótulo inicial do texto. Ao mesmo tempo, determinar se um idioma é mais simples ou complexo que outro ainda é uma questão controversa e dependente do contexto; inclusive, estudos recentes apontam que não é possível medir se uma língua é mais complexa que outra (BENTZ *et al*., 2022).

Assim, para a escolha dos idiomas intermediários envolvidos na tarefa, foi considerado o fato de que as línguas de família românica, como o francês, espanhol, italiano e o português, tendem a ter um espectro maior de variáveis no plano sintático, em termos de gênero gramatical, declinação de substantivos e adjetivos, bem como conjugações verbais, enquanto as línguas anglo-saxônicas, como o inglês, o alemão e o holandês, tendem a ser mais simples nesse sentido (PEI; GAYNOR, 1954). Por sua vez, o italiano pode ser considerado o idioma que conserva mais palavras com maior proximidade das palavras latinas, logo com um maior número de declinações; e o inglês, além de ser um idioma com elevado número de recursos computacionais, é conhecido por ter poucas modificações na conjugação de verbos e inexistência de gêneros gramaticais.

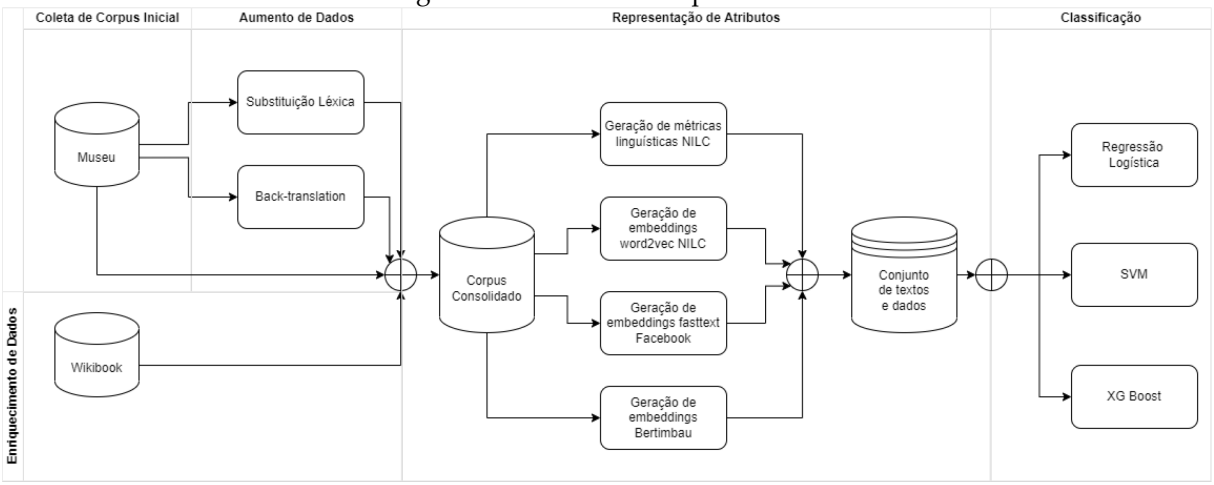
Nesse contexto de especificidades idiomáticas e gramaticais envolvidas, realizou-se a varredura nos textos do *corpus* de treinamento, considerando seus respectivos rótulos originais para identificação da língua intermediária. Assim, optou-

se pela retrotradução com a língua inglesa como intermediária dos textos simples, e da língua italiana para os textos complexos.

4.2 Metodologia Experimental

A abordagem proposta neste trabalho desenvolve uma análise embutida de mecanismos de AD em termos de leiturabilidade para o PB. Neste contexto, definem-se cinco etapas principais, incluindo a aplicação dos métodos de aumento de dados delineados na seção anterior, representadas na Figura 1. As demais quatro etapas são descritas, em mais detalhes, na sequência.

Figura 1 – Desenho do experimento.



Fonte: elaboração própria .

4.3 Coleta de *Corpora*

***Corpus* Alvo:** As análises aqui apresentadas concebiam um conjunto reduzido de textos como alvo principal, anotado e pareado por uma equipe de linguistas liderada na UFRGS (FINATTO; TCACENCO, 2021). O *corpus* em questão está representado na Figura 1 pelo conjunto *Museu* e contém 42 textos originais e suas simplificações, totalizando 84 textos. São textos cuja função é acompanhar experimentos expostos em um museu gaúcho de ciências e tecnologia. Os textos originais foram escritos para público leigo em geral, e suas versões simplificadas foram

adaptadas lexical e sintaticamente para alunos do final do Ensino Fundamental de escolas públicas por especialistas (FINATTO; TCACENCO, 2021).

Enriquecimento de Dados para Teste: para avaliar se os resultados obtidos a partir das técnicas de aumento de dados poderiam ser generalizados para outro domínio, buscaram-se dados de textos originais e simplificados pareados de outros domínios do conhecimento para teste. Dentre os trabalhos identificados na Seção 3, o único repositório desse gênero ativo e disponível foi o disponibilizado por Wilkens *et al*. (2016). Esse trabalho proporcionou a consulta ao *corpus* Wikilivros, previamente anotado em relação ao sistema de educação brasileiro. Para conversão em "simples" e "complexo", considerou-se apenas que o primeiro nível (Ensino Fundamental) conteria os textos simples, e os demais, complexos. Os textos coletados via *web*, por Wilkens *et al*. (2016), foram desconsiderados, dada a possibilidade de imprecisão da anotação disponibilizada, visto terem sido rotulados por meio de um modelo automático de classificação, o que tenderia à propagação de erros.

Sobre os textos da camada de enriquecimento, devido à limitação da ferramenta NILC-Metrix de 2000 palavras para processamento e do tempo para processamento de *embeddings* para textos muito longos, foram consideradas apenas as primeiras 2000 palavras dos textos para captura da representação de atributos.

4.4 Classificação

Para análise dos conjuntos gerados artificialmente, além de uma ponderação holística em termos qualitativos, considerou-se a tarefa de AAL. O trabalho de Wilkens *et al*. (2016) é particularmente interessante, haja vista que o problema de pesquisa se aproxima do trabalho aqui apresentado. Neste contexto, para fins comparativos, adota-se aqui também o uso de um RL para o treinamento do classificador automático.

Devido ao baixo número de textos do *corpus* alvo, emprega-se o método conhecido por *leave one-out (LOO)*[12] para treinamento e validação dos resultados do classificador. Desta forma, considerando o *corpus* inicial com i = 84 textos, o modelo foi treinado 84 vezes, sendo que cada vez um texto específico era desconsiderado do conjunto de treinamento e usado apenas para avaliar o modelo. Ressalta-se que além do texto em si, o seu par anotado com etiqueta oposta também foi desconsiderado nos treinos, de modo a evitar um vazamento dos textos de teste no treinamento e ocasionar sobre ajustes no classificador. No caso dos treinos com uso de AD, os textos aumentados relacionados ao texto excluído, bem como ao seu par oposto também não são considerados naquela rodada.

4.5 Representação de Atributos

Neste trabalho, explora-se tanto o uso das representações por incorporação de palavras quanto dos conjuntos de métricas de análise linguística e psicolinguística disponíveis para o PB gerados pelo NILC-Metrix (LEAL *et al*., 2021; LEAL *et al*., 2022a), considerando análises combinatórias de ambos os formatos de representação de atributos.

Para a extração das incorporações de palavras, são consideradas abordagens estáticas e contextualizadas. No caso das estáticas, foram utilizadas as bibliotecas: (1) Fasttext (INC., 2022) com carga do modelo pré-treinado *cc.pt.300.bin* (Inc., 2018), que faz uso da abordagem *word2vec* por *CBOW*, e (2) *Gensim* (REHUREK, 2022) com carga do modelo pré treinado *skip s300* (NILC, 2017), que faz uso de *word2vec* por *skip gram*. Em ambos os modelos, utilizamos *embeddings* com 300 dimensões.

[12] O LOO é um caso especial da validação cruzada conhecida por *k-fold*, em que o número de divisões do conjunto de treinamento k é igual ao número de instâncias i do conjunto de dados. De acordo com Wong (2015) é indicado o uso do LOO para conjuntos de dados pequenos para se obter uma estimativa de precisão mais confiável em algoritmos de classificação.

Em relação às abordagens contextualizadas, foi utilizada a biblioteca *Transformers*, disponibilizada pelo grupo *Hugging Face* (WOLF *et al*., 2020) com carga do modelo pré-treinado BERTimbau disponibilizado pelo nome de *neuralmind/bert-base portuguese-cased* (SOUZA *et al*., 2020) com 768 dimensões.

Especificamente sobre o modelo BERTimbau, há uma restrição de processamento no número de *tokens*, de modo que não é possível converter uma sequência maior do que 512 *tokens* de uma vez. Por isso, foi utilizada uma estratégia de processamento em que o texto é quebrado em sentenças de até X palavras separadas de forma gráfica. Ao final do processamento das sentenças, é realizada uma média aritmética entre os vetores gerados para representar cada um dos textos do *corpus*.

5 Resultados experimentais

Para avaliar extrinsecamente as técnicas de AD propostas neste trabalho, foram consideradas 75 combinações de métodos e conjuntos de textos para treinamento de um modelo supervisionado de classificação por leiturabilidade (Tabela 1), voltados para AAL.

Observa-se, na Tabela 1, que o limiar que define os agrupamentos dos modelos desenvolvidos é altamente variável, de modo que pequenas alterações na seleção de atributos ou dos textos utilizados para treinamento impactam significativamente o resultado. Ressalta-se que, devido ao baixo número de textos considerados neste experimento, o viés de tal análise tem relação direta com a volumetria dos dados.

Em termos holísticos, verificou-se que os textos aumentados por RT apresentavam maior nível de manutenção da estrutura gramatical do que aqueles aumentados por SS. Por outro lado, no que diz respeito às avaliações quantitativas, destaca-se que o aumento por SS promoveu melhores resultados que o RT, mas não superiores à linha de base dada pelo modelo não aumentado. Desta forma, o melhor resultado obtido para o conjunto inicial dos textos do *corpus* Museu foi de 94,0% considerando um modelo de RL com entrada apenas por *embeddings* contextualizados.

Este resultado foi ligeiramente melhorado ao combinar as métricas do NILC-Metrix com os *embeddings* contextualizados, para o modelo treinado com aumento por RT para a classe simples e SS para ambas as classes.

Destaca-se que, os modelos treinados com uso de dados aumentados demonstraram uma capacidade igual ou levemente superior àqueles treinados sem aumento e, ao mesmo tempo, apresentaram maior generalização quando introduzidos a outros domínios linguísticos, como pode ser visto nos resultados com os *corpora* de Wilkens *et al.* (2016).

Adicionalmente, em termos comparativos com trabalhos anteriores, ainda que não seja possível estabelecer uma relação direta entre o trabalho atual e o de Imperial (2021) haja vista que os *corpora* utilizados diferem entre si, numa análise genérica de textos em idiomas de menos recursos, os resultados obtidos foram significativamente superiores aos indicados por Imperial (2021) para a língua filipina. Comparando com resultados no PB, foi possível validar a generalização do modelo treinado com os textos do *corpus* Museu ao testar o classificador com o mesmo *corpus* utilizado e desenvolvido no trabalho de Wilkens *et al*. (2016), o Wikilivros. O uso da combinação de textos aumentados, *embeddings* do BERTimbau e métricas do NILC-Metrix, aqui proposto, atingiu 84,4% de acurácia para o classificador, superior à acurácia do classificador treinado sem o aumento. Assim sendo, podemos afirmar que o treino do classificador por meio de um conjunto reduzido avaliado por especialistas em leiturabilidade e aumentados por técnicas simples de AD, promoveu resultados mais controlados quando comparados ao uso de técnicas massivas como a *WaC* proposta em Wilkens *et al*. (2016).

Tabela 1 — Resultados das Classificações considerando diferentes combinações de *corpora*, atributos e técnicas classificatórias.

***Corpus* de Teste**	***Corpus* de Treino**	**Técnica de AD**	**ML[1]**	**Embeddings**	**CLF[2]**	**Tipo de Validação**	**ACC[3] (%)**
Museu	SS_Museu, RT_Museu (classe simples), Museu	RT e SS	✓	Apenas BERTimbau	RL	LOO	95,20
Museu	Museu	✕	✕	Apenas BERTimbau	RL	LOO	94,00
Museu	Museu	✕	✓	Apenas BERTimbau	RL	LOO	94,00
Museu	SS_Museu (classe complexa), RT_Museu (classe simples), Museu	RT e SS	✓	Apenas BERTimbau	RL	LOO	94,00
Museu	SS_Museu, Museu	SS	✓	Apenas BERTimbau	RL	LOO	94,00
Museu	SS_Museu, Museu	SS	✓	Apenas BERTimbau	RL	LOO	94,00
Museu	SS_Museu, Museu	SS	✓	Apenas BERTimbau	RL	LOO	94,00
Museu	RT_Museu (classe simples), Museu	RT	✓	Apenas BERTimbau	RL	LOO	92,90
Museu	Museu	✕	✓	✓	SVM[4]	LOO	92,90

[1] **Métricas Linguísticas**
[2] **Classificador**
[3] **Acurácia**
[4] Sigla para *Support Vector Machine*, ou máquina de vetores de suporte em português. É uma técnica de AM para classificação por meio de um modelo que se baseia em encontrar um hiperplano capaz de separar o conjunto de dados em classes distintas.

Museu	SS_Museu, BT_Museu, Museu	RT e SS	✕	Apenas BERTimbau	RL	LOO	92,90
Museu	BT_Museu, Museu	RT	✕	Apenas BERTimbau	RL	LOO	91,70
Museu	Museu	✕	✕	Apenas BERTimbau	SVM	LOO	91,70
Museu	Museu	✕	✓	✓	RL	LOO	91,70
Museu	SS_Museu (classe complexa), RT_Museu (classe simples), Museu	RT e SS	✕	Apenas BERTimbau	RL	LOO	91,70
Museu	SS_Museu, Museu	SS	✕	Apenas BERTimbau	RL	LOO	91,70
Museu	SS_Museu, Museu	SS	✓	Apenas BERTimbau	SVM	LOO	91,70
Museu	SS_Museu, Museu	SS	✕	Apenas BERTimbau	RL	LOO	91,70
Museu	SS_Museu, Museu	SS	✕	Apenas BERTimbau	RL	LOO	91,70
Museu	BT_Museu, Museu	RT	✓	Apenas BERTimbau	RL	LOO	90,50
Museu	BT_Museu, Museu	RT	✓	Apenas BERTimbau	RL	LOO	90,50
Museu	SS_Museu, BT_Museu, Museu	RT e SS	✓	Apenas BERTimbau	RL	LOO	90,50
Museu	SS_Museu (classe complexa), RT_Museu (classe complexa), Museu	RT e SS	✓	Apenas BERTimbau	RL	LOO	90,50
Museu	SS_Museu (classe simples), RT_Museu (classe complexa), Museu	RT e SS	✓	Apenas BERTimbau	RL	LOO	90,50

Museu	SS_Museu (classe simples), RT_Museu (classe simples), Museu	RT e SS	✕	Apenas BERTimbau	RL	LOO	90,50
Museu	RT_Museu (classe simples), Museu	RT	✕	Apenas BERTimbau	RL	LOO	89,30
Museu	SS_Museu, RT_Museu (classe simples), Museu	RT e SS	✕	Apenas BERTimbau	RL	LOO	89,30
Museu	SS_Museu (classe complexa), RT_Museu (classe complexa), Museu	RT e SS	✕	Apenas BERTimbau	RL	LOO	88,10
Museu	SS_Museu (classe simples), RT_Museu (classe complexa), Museu	RT e SS	✕	Apenas BERTimbau	RL	LOO	88,10
Museu	BT_Museu, Museu	RT	✕	Apenas BERTimbau	RL	LOO	86,90
Museu	Museu	✕	✕	Apenas Facebook CBOW	RL	LOO	85,70
Museu	Museu	✕	✓	✓	XGB[5]	LOO	85,70
Museu	Museu	✕	✓	✕	RL	LOO	84,50
Museu	Museu	✕	✓	✕	SVM	LOO	84,50
Museu	Museu	✕	✓	✕	XGB	LOO	84,50
Museu	SS_Museu, Museu	SS	✓	✕	RL	LOO	84,50

[5] Sigla para *Extreme Gradient Boosting*, ou aumento extremo de gradiente em português. É uma técnica de AM para problemas de classificação e regressão, que considera o uso de árvores de decisão em um modelo de retroalimentação.

Museu	Museu	✕	✕	Apenas BERTimbau	XGB	LOO	79,80
Museu	Museu	✕	✕	Apenas NILC Skip Gram	RL	LOO	79,80
Museu	Museu	✕	✕	Apenas Facebook CBOW	SVM	LOO	76,20
Museu	Museu	✕	✕	Apenas Facebook CBOW	XGB	LOO	76,20
Museu	Museu	✕	✕	Apenas NILC Skip Gram	SVM	LOO	73,80
Museu	Museu	✕	✕	Apenas NILC Skip Gram	XGB	LOO	71,40
Wikibooks	SS_Museu, BT_Museu, Museu	RT e SS	✓	Apenas BERTimbau	RL	Por seleção	84,40
Wikibooks	Museu	✕	✓	Apenas BERTimbau	RL	Por seleção	83,10
Wikibooks	SS_Museu, RT_Museu (classe simples), Museu	RT e SS	✓	Apenas BERTimbau	RL	Por seleção	83,10
Wikibooks	RT_Museu (classe simples), Museu	RT	✓	Apenas BERTimbau	RL	Por seleção	80,50
Wikibooks	SS_Museu, BT_Museu, Museu	RT e SS	✕	Apenas BERTimbau	RL	Por seleção	80,50
Wikibooks	SS_Museu (classe complexa), RT_Museu (classe simples), Museu	RT e SS	✓	Apenas BERTimbau	RL	Por seleção	80,50
Wikibooks	Museu	✕	✕	Apenas BERTimbau	RL	Por seleção	79,20
Wikibooks	SS_Museu, RT_Museu (classe simples), Museu	RT e SS	✕	Apenas BERTimbau	RL	Por seleção	79,20
Wikibooks	SS_Museu, Museu	SS	✓	Apenas BERTimbau	RL	Por seleção	79,20

Wikibooks	SS_Museu (classe simples), Museu	SS	✓	Apenas BERTimbau	RL	Por seleção	79,20
Wikibooks	RT_Museu (classe simples), Museu	RT	✕	Apenas BERTimbau	RL	Por seleção	76,60
Wikibooks	SS_Museu (classe complexa), RT_Museu (classe simples), Museu	RT e SS	✕	Apenas BERTimbau	RL	Por seleção	76,60
Wikibooks	SS_Museu, Museu	SS	✕	Apenas BERTimbau	RL	Por seleção	75,30
Wikibooks	SS_Museu (classe simples), RT_Museu (classe simples), Museu	RT e SS	✕	Apenas BERTimbau	RL	Por seleção	72,70
Wikibooks	SS_Museu (classe simples), RT_Museu (classe simples), Museu	RT e SS	✓	Apenas BERTimbau	RL	Por seleção	71,40
Wikibooks	SS_Museu (classe simples), Museu	SS	✕	Apenas BERTimbau	RL	Por seleção	70,10
Wikibooks	Museu	✕	✓	Apenas BERTimbau	RL	Por seleção	65,50
Wikibooks	Museu	✕	✕	Apenas BERTimbau	RL	Por seleção	64,30

Fonte: elaboração própria.

6 Considerações finais

Por meio desta análise exploratória, observou-se que a assertividade do modelo de classificação automática para auxiliar a tarefa de avaliação de leiturabilidade está diretamente relacionada ao conjunto de textos escolhido para treino. É interessante que o *corpus* contenha textos dentro de um espectro de variações dos padrões lexicais e sintáticos para maior generalização, o que pode ser introduzido via métodos de AD. Neste sentido, trabalhos futuros devem considerar a inclusão de outros textos no conjunto de treinamento, garantindo os níveis de complexidade textual do *corpus* e possibilitando a utilização de técnicas como *fine-tuning* com menor chance de *overfitting*.

Ainda que as possibilidades de combinações e de uso de técnicas de AD sejam diversas, a inclusão de textos aumentados proporcionou uma maior generalização do modelo ao considerar diferentes domínios, tal que o resultado da combinação das tarefas de SS e RT promoveram resultados superiores do que quando usadas individualmente. Além disso, este estudo permitiu corroborar, para o PB, o conceito de que a leiturabilidade está fortemente atrelada ao processo de predição probabilística dada por uma vizinhança de palavras, conforme supunha Taylor (1953), visto que os classificadores com melhores resultados foram aqueles treinados com uso de *embeddings* contextualizados. Tal fato valida a argumentação de Imperial (2021) de que o conhecimento implicitamente codificado nos *embeddings* contextualizados pode ser usado como um conjunto para idiomas com recursos de dados com baixa volumetria, no que tange à leiturabilidade.

Agradecimentos

Agradecemos o suporte financeiro do CNPq (Proc. 311275/2020-6 e Proc. 308926/2019-6 e da FAPERJ (processos SEI-260003/000614/2023 e E-26/202.914/2019). Obrigada também pelo trabalho dos revisores, com comentários e sugestões, e dos editores, pelo encaminhamento do artigo.

Referências

AGGARWAL, C. C. **Machine learning for text,** v. 848, Springer, 2018. DOI https://doi.org/10.1007/978-3-319-73531-3

ALUÍSIO, S. M.; SPECIA, L.; PARDO, T. A. S.; MAZIERO, E. G.; FORTES, R. Towards Brazilian Portuguese automatic text simplification systems. **Proceedings of the eighth ACM symposium on Document engineering**, 2008. p. 240–248. DOI https://doi.org/10.1145/1410140.1410191

ALUÍSIO, S.; SPECIA, L.; GASPERIN, G.; SCARTON, C. Readability assessment for text simplification. **Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications**, 2010. p. 1–9

BACCOURI, N. deep-translator. **Deep translator**. 2020. Disponível em: https://deep-translator.readthedocs.io/en/latest/README.html. Acesso em : 20 fev. 2023.

BAYER, M.; KAUFHOLD, M.; REUTER, C. A survey on data augmentation for text classification. **ACM Computing Surveys**, 2021. DOI https://doi.org/10.1145/3544558

BENTZ, C. *et al*. Complexity Trade-Offs and equi-complexity in natural languages: A meta-analysis. **Linguistics Vanguard**, De Gruyter Mouton, 2022. DOI https://doi.org/10.1515/lingvan-2021-0054

BIRD, S.; KLEIN, E.; LOPER, E. Natural language processing with python: analyzing text with the natural language toolkit. **O'Reilly Media, Inc**, 2019.

BRILL, E. Processing natural language without natural language processing. **International Conference on Intelligent Text Processing and Computational Linguistics**, Springer, 2003. p. 360–369. DOI https://doi.org/10.1007/3-540-36456-0_37

CAO, Y.; SHUI, R.; PAN, L.; KAN, M. Y.; LIU, Z.; CHUA, T. Expertise style transfer: A new task towards better communication between experts and laymen. **58th Annual Meeting of the Association for Computational Linguistics**, Online: Association for Computational Linguistics, 2020. p. 1061–1071. DOI https://doi.org/10.18653/v1/2020.acl-main.100

CAYLOR, J. S. *et al*. Methodologies for determining reading requirements of military occupational specialties, **ERIC**, 1973.

CHAQUET-ULLDEMOLINS, J.; GIMENO-BLANES, F.; MORAL-RUBIO, S.; MUNOZ-ROMERO, S.; ROJO-ALVAREZ, J. L. On the black-box challenge for fraud detection using machine learning (ii): Nonlinear analysis through interpretable

autoencoders. **Applied Sciences.** v.12(8), p. 3856, 2022. DOI https://doi.org/10.3390/app12083856

CHUCHU, M. **Readability Assessment with Pre-Trained Transformer Models**: An Investigation with Neural Linguistic Features. Uppsala University, 2022.

DEVLIN, J.; CHANG, M. W.; LEE, K.; TOUTANOVA, K. Bert: Pre training of deep bidirectional transformers for language understanding. **Proceedings of naacL-HLT**, v. 1, 2018. p. 2.

DIAS-DA-SILVA, B. C.; MORAES, H. R. A construção de um thesaurus eletrônico para o português do Brasil. **ALFA**: Revista de Linguística, 2003.

FENG, L.; ELHADAD, N.; HUENER, M. Cognitively motivated features for readability assessment. **12th Conference of the European Chapter of the ACL**, 2009. p. 229–237. DOI https://doi.org/10.3115/1609067.1609092

FENG, S. Y.; GANGAL, V.; WEI, J.; CHANDAR, S.; VOSOUGHI, S.; MITAMURA, T.; HOVY, E. 2021. A survey of data augmentation approaches for nlp. **Association for Computational Linguistics**, 2021. p. 968-988. DOI https://doi.org/10.18653/v1/2021.findings-acl.84

FINATTO, M. J. B. 2020. Acessibilidade textual e terminológica: promovendo a tradução intralinguística. **Estudos Linguísticos.** São Paulo, v. 49(1), p. 72–96, 2020. DOI https://doi.org/10.21165/el.v49i1.2775

FINATTO, M. J. B; TCACENCO, L. M. Tradução intralinguística, estratégias de equivalência e acessibilidade textual e terminológica. **Tradterm**. São Paulo, v. 37, n. 1, p. 30-63, 2021. DOI https://doi.org/10.11606/issn.2317-9511.v37p30-63

FREITAS, C. Linguística computacional. **Parábola Editorial**, 2022.

HARTMANN, N. S.; ALUÍSIO, S. M. Adaptação lexical automática em textos informativos do português brasileiro para o ensino fundamental. **Linguamática,** v.12(2). p. 3–27, 2020. DOI https://doi.org/10.21814/lm.12.2.323

HONNIBAL, M.; MONTANI, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. **To appear**. 2017.

IMPERIAL, J. M. BERT embeddings for automatic readability assessment. *In*: **Proceedings of the International Conference on Recent Advances in Natural Language Processing**, v.1-3, INCOMA Ltd, 2021. p. 611–618. Disponível em: https://aclanthology.org/2021.ranlp-1.69. Acesso em: 20 fev. 2023.

INC., F. Learning word vectors for 157 languages. **Fasttext**. 2018. Disponível em: https://dl.fbaipublicfiles.com/fasttext/vectorscrawl/cc.pt.300.bin.gz. Acesso em: 20 fev. 2023.

INC., F. Library for efficient text classification and representation learning. **Fasttext**. 2022. Disponível em: https://fasttext.cc/. Acesso em: 20 fev. 2023.

KINCAID, J. P.; YASUTAKE, J. Y.; GEISELHART, R. Use of the automated readability index to assess comprehensibility of air force technical orders. **Wright-Patterson AFB**, Ohio: Aeronautical Systems Division, 1967.

KLARE, G. R. *et al*. Measurement of readability, **AGRIS**, Iowa State University Press, 1963.

LEAL, S. E. 2020. Simpligo ranking. **NILC**. Disponível em: http://fw.nilc.icmc.usp.br:23380/simpligo-ranking. Acesso em: 20 fev. 2023.

LEAL, S. E. **Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular**. 2019. Tese de Doutoramento. Universidade de São Paulo, 2019.

LEAL, S. E.; DURAN, M. S.; SCARTON, C. E.; HARTMANN, N. S.; ALUISIO, S. M. Nilc Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. **CoRR**, v.2201.03445, 2021.

LEAL, S. E.; LUKASOVA, K.; CARTHERY-GOULART, M. T.; ALUISIO, S. M. Rastros project: Natural language processing contributions to the development of an eye-tracking corpus with predictability norms for Brazilian Portuguese. **Language Resources and Evaluation,** v.56(4), p.1333– 1372, 2022b. DOI https://doi.org/10.1007/s10579-022-09609-0

LEAL, S. E.; MAGALHAES, V. M. A.; DURAN, M. S.; ALUÍSIO, S. M. Avaliação automática da complexidade de sentenças do português brasileiro para o domínio rural. **Symposium in Information and Human Language Technology and Collocates**, 2019.

LEAL, S. E.; SCARTON, C.; CUNHA, A.; HARTMANN, N.; DURAN, M.; ALUÍSIO, S. Nilc-Metrix, **NILC**, 2022a. Disponível em: http://fw.nilc.icmc.usp.br:23380/nilcmetrix. Acesso em: 20 fev. 2023.

LEAL, S. E; DURAN, M. S.; ALUÍSIO, S. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. **27th International Conference on Computational Linguistics**, 2018. p. 401–413.

LEE, B. W.; JANG, Y. S.; LEE, J. H. J. Pushing on text readability assessment: A transformer meets hand crafted linguistic features, **CoRR**, v.2109.12258, 2021. DOI https://doi.org/10.18653/v1/2021.emnlp-main.834

LONGPRE, S., WANG, Y.; BOIS, C. D. How effective is task-agnostic data augmentation for pretrained transformers? **Findings of ACL**, v. EMNLP 2020, 2020. p. 4401-4411. DOI https://doi.org/10.18653/v1/2020.findings-emnlp.394

MARTINC, M.; POLLAK, S.; SIKONJA, M. R. Supervised and Unsupervised Neural Approaches to Text Readability. **Computational Linguistics,** v.47(1), p. 141–179, 2021. DOI https://doi.org/10.1162/coli_a_00398

MARTINS, T. B. F.; GHIRALDELO, C. M.; NUNES, MGV; JUNIOR, ONO. Readability formulas applied to textbooks in Brazilian Portuguese, **Icmsc-Usp**, 1996.

MCNAMARA, M. M.; LOUWERSE, D. S.; GRAESSER, A. C. Coh-metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. **Cognitive Science and Educational Practice group at the University of Memphis**, 2002. Disponível em: http://csep.psyc.memphis.edu/mcnamara/pdf/IESproposal.pdf. Acesso em: 20 fev. 2023.

MUNIZ, M. C. M. **A construção de recursos linguístico-computacionais para o português do Brasil**: o projeto unitex-pb. Tese de Doutoramento. Universidade de São Paulo, 2004.

NILC. Repositório de word embeddings. **NILC**. Disponível em: http://143.107.183.175:22980/download.php?file=embeddings/fasttext/skip_s300.zip. 2017. Acesso em : 20 fev. 2023.

PASQUALINI, B. F. **CorPop**: um corpus de referência do português popular escrito do Brasil. Tese de Doutoramento. Universidade Federal do Rio Grande do Sul, 2018.

PEI, M.; GAYNOR, F. Dictionary of linguistics. **Rowman & Littlefield**, 1954.

PONOMARENKO, G. L. Índices para cálculo de leiturabilidade. **Acessibilidade TT**, Universidade Federal do Rio Grande do Sul, 2018. Disponível em: http://www.ufrgs.br/textecc/acessibilidadett/files/IndicesdeLeiturabilidade.pdf. Acesso em : 20 fev. 2023.

REHUREK, R. Gensim topic modelling for humans. **Gensim**, 2022. Disponível em: https://radimrehurek.com/gensim/index.html. Acesso em: 20 fev. 2023.

SAHIN, G. G. To augment or not to augment? a comparative study on text augmentation techniques for low resource NLP. **Computational Linguistics**, v. 48(1), p. 5–42, 2022. DOI https://doi.org/10.1162/coli_a_00425

SANTOS, A. M. Leiturabilidade: É possível medi-la em livros infanto-juvenis? **Congresso Internacional de Leitura e Literatura Infantil e Juvenil**, 2010.

SCARTON, C.; ALUÍSIO, S. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh metrix para o português. **Linguamática,** v. 2(1). p.45–61, 2010.

SCARTON, C.; GASPERIN, C.; ALUÍSIO, S. Revisiting the readability assessment of texts in Portuguese. **Ibero-American Conference on Artificial Intelligence**, Springer, 2010a. p. 306–315. DOI https://doi.org/10.1007/978-3-642-16952-6_31

SCARTON, C.; OLIVEIRA, M.; JÚNIOR, A. C.; GASPERIN, C.; ALUÍSIO, S. Simplifica: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. **Proceedings of the NAACL HLT 2010 Demonstration Session**, 2010b. p. 41–44.

SCHWARM, S.; OSTENDORF, M. Reading level assessment using support vector machines and statistical language models. **43rd annual meeting of the Association for Computational Linguistics**, 2005. p. 523–530. DOI https://doi.org/10.3115/1219840.1219905

SHUKLA, A.; PANDEY, H. M.; MEHROTRA, D. Comparative review of selection techniques in genetic algorithms. **International conference on futuristic trends on computational analysis and knowledge management,** IEEE, 2015. p. 515–519. DOI https://doi.org/10.1109/ABLAZE.2015.7154916

SI, L.; CALLAN, J. A statistical model for scientific readability. **International conference on Information and knowledge management**, 2001. p. 574–576. DOI https://doi.org/10.1145/502585.502695

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. **9th Brazilian Conference on Intelligent Systems**, 2020. DOI https://doi.org/10.1007/978-3-030-61377-8_28

STOLCKE, A. Srilm - an extensible language modeling toolkit. **Seventh international conference on spoken language processing**, ISCA, 2002. DOI https://doi.org/10.21437/ICSLP.2002-303

TAYLOR, W. L. "cloze procedure": A new tool for measuring readability. **Journalism quarterly,** v.30(4), 1953. p. 415–433. DOI https://doi.org/10.1177/107769905303000401

TAYNAN, M. F.; COSTA, A. H. R. Deepbt and nlp data augmentation techniques: A new proposal and a comprehensive study. **Ricardo Cerri & Ronaldo C. Prati**, Intelligent Systems, Springer, 2020. p. 435– 449. DOI https://doi.org/10.1007/978-3-030-61377-8_30

USP. A lexicon for Brazilian Portuguese according to Universal Dependencies. **PortiLexicon-UD**. 2022a. Disponível em: https://portilexicon.icmc.usp.br/. Acesso em: 20 fev. 2023.

USP. **Tep 2.0**. 2022b. Disponível em: http://www.nilc.icmc.usp.br/tep2/ajuda.htm. Acesso em: 20 fev. 2023.

USP; FAPESP; IBM. Portuguese processing - towards syntactic analysis and parsing. **Poetisa**. 2021. Disponível em: https://sites.google.com/icmc.usp.br/poetisa. Acesso em: 20 fev. 2023.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZJOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is all you need. **Advances in neural information processing systems,** p.5998-6008, 2017.

WATANABE, W. M.; JUNIOR, A. C.; UZÊDA, V. R.; FORTES, R. P. M.; PARDO, T. A. S.; ALUÍSIO, S. M. Facilita: reading assistance for low literacy readers. **Proceedings of the 27th ACM international conference on Design of communication**, 2009. p. 29–36. DOI https://doi.org/10.1145/1621995.1622002

WILKENS, R.; VECCHIA, A. D.; BOITO, M. Z.; PADRÓ, M.; VILLAVICENCIO, A. Size does not matter. frequency does. a study of features for measuring lexical complexity. **Ibero-American conference on artificial intelligence**, Springer, 2014. p. 129–140. DOI https://doi.org/10.1007/978-3-319-12027-0_11

WILKENS, R.; ZILIO, L.; IDIART, M.; VILLAVICENCIO, A. *et al*. Crawling by readability level. **International Conference on Computational Processing of the Portuguese Language**, Springer, 2016. p. 306–318. DOI https://doi.org/10.1007/978-3-319-41552-9_31

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J. *et al*. Transformers: State-of-the-art natural language processing. **Association for Computational Linguistics**, 2020. p. 38-45. DOI https://doi.org/10.18653/v1/2020.emnlp-demos.6

WONG, T. T. Performance evaluation of classification algorithms by k-fold and leave one-out cross validation. **Pattern Recognition**, v. 48(9), p. 2839–2846, 2015. DOI https://doi.org/10.1016/j.patcog.2015.03.009

XIA, M.; KOCHMAR, E.; BRIS, B. Text readability assessment for second language learners. **The Association for Computer Linguistics**, p. 12-22, 2019.